Supplementary Material for the paper "Asymptotics for Constrained Dirichlet Distributions"

By

Charles J. Geyer and Glen D. Meeden

Technical Report No. 691

School of Statistics
University of Minnesota

March 20, 2012

Abstract

This document is supplementary material for a paper. It shows how to simulate the linear-equality-and-inequality-constrained normal distribution that is the large sample approximation to a similarly constrained Dirichlet posterior.

1 Introduction

Throughout this document "the paper" refers to Geyer and Meeden (submitted). We are interested in simulating a constrained Dirichlet distribution that is a posterior for multinomial data using conjugate priors. According to the discussion in Section 4.2 of the paper, if y is the vector of multinomial data, then the unconstrained posterior is Dirichlet with hyperparameter $\hat{\alpha}_n = y + \xi$, where ξ is the hyperparameter of the prior (which is also Dirichlet). Here we use $\xi = 0$, which gives an improper prior but a proper posterior so long as no component of y is zero. The constrained Dirichlet posterior restricts the Dirichlet probability density function to a constraint set determined by a finite set of linear equality and inequality constraints and renormalizes so it integrates to one over the constraint set.

According to Remark 6 in the paper, we approximate the posterior with a normal distribution having unnormalized probability density function

$$h(\lambda) = \exp\left(-\frac{n}{2}(\lambda - \hat{\lambda}_n)^T \widehat{\Lambda}_n^{-1} (\lambda - \hat{\lambda}_n)\right) \tag{1}$$

(this is equation (22) in the paper), where the scalar n is the multinomial sample size, the vector $\hat{\lambda}_n$ is the unconstrained posterior mean $\hat{\alpha}_n/n$, and the matrix $\widehat{\Lambda}_n$ is diagonal with diagonal entries that are the corresponding components of the vector $\hat{\lambda}_n$. The constrained asymptotic approximation of the posterior restricts (1) to the constraint set and renormalizes so it integrates to one over the constraint set.

We impose the constraints in two steps: equality constraints first, and then inequality constraints. The equality constraints constrain the parameter vector λ to an affine subspace.

Suppose the equality constraints have the form $B\lambda = a$ where B is a known matrix and a is a known vector, let V denote the vector subspace

$$V = \{\, w \in \mathbb{R}^d : Bw = 0 \,\}$$

(this is equation (19) in the paper), and let M be a matrix whose columns are a basis for V. Then every λ in the equality constraint set has the form

$$\lambda = \lambda_0 + M\beta \tag{2}$$

for some β, where λ_0 is any point in the equality constraint set, that is, any vector satisfying $B\lambda_0 = a$. We can think of β as a new parameter for the problem that, like λ has an approximate normal distribution, and, moreover, unlike λ, has a nondegenerate normal approximation. The mean and variance of this normal approximation are given in Remark 6 of the paper as

$$\beta_n^* = \left(M^T \widehat{\Lambda}_n^{-1} M\right)^{-1} M^T \widehat{\Lambda}_n^{-1} (\hat{\lambda}_n - \lambda_0) \tag{3}$$

(this is equation (25) in the paper) and

$$n^{-1}(M^T \widehat{\Lambda}_n^{-1} M)^{-1} \tag{4}$$

We simulate λ having the normal approximation to the equality constrained Dirichlet posterior by simulating β multivariate normal with mean vector (3) and variance matrix (4) and then transforming to the original parameter λ via (2).

We then impose the inequality constraints. In the simulation we reject any λ that do not satisfy the constraints.

2 R Package RCDD

We use the R statistical computing environment (R Development Core Team, 2012) in our analysis. It is free software and can be obtained from `http://cran.r-project.org`. Precompiled binaries are available for Windows, Macintosh, and popular Linux distributions. We use the contributed package `rcdd` (Geyer and Meeden, 2009). If R has been installed, but this package has not yet been installed, do

```
install.packages("rcdd")
```

from the R command line (or do the equivalent using the GUI menus if on Apple Macintosh or Microsoft Windows). This may require root or administrator privileges.

Assuming the `rcdd` package has been installed, we load it

```
> suppressPackageStartupMessages(library(rcdd))
```

The version of the package used to make this document is 1.1-7. The version of R used to make this document is 2.15.0.

This entire document and all of the calculations shown were made using the R command `Sweave` and hence are exactly reproducible by anyone who has R and the R noweb (RNW) file from which it was created. Both the RNW file and and the PDF document produced from it will be made available at the University of Minnesota Digital Conservancy.

Not only can one exactly reproduce the results in the printable document, one can also modify the parameters of the simulation and get different results. Anything at all can be changed once one has the RNW file.

In particular, we set the "seed" of the random number generator

```
> set.seed(42)
```

so that every time this RNW file is run it produces the same results. Changing the argument of `set.seed` or removing this chunk of R code will produce different results.

```
> d <- 9
> n <- 1000
> if (d < 7) stop("need dimension at least 7")
```

These statements choose the dimension of the parameter space (9) and the multinomial sample size (1000). Modifying either or both of these statements will change the simulation accordingly.

3 The Constraints

We set up the constraints as an RCDD H-representation. First, the constraints defining the unit simplex.

```
> hrep <- makeH(- diag(d), rep(0, d), rep(1, d), 1)
```

These are the constraints that the components of the parameter are nonnegative and sum to one.

Second, a constraint on the mean of a certain random variable. Let X be a random variable taking values $1, \ldots, d$ with probabilities given by the vector λ.

```
> x <- d2q(1:d)
> mu <- qdq(qsum(x), d2q(d))
> hrep <- addHeq(x, mu, hrep)
```

This equality constraint requires that the mean of X be $\mu = (d+1)/2 = 5$.

Third, constraints on the median of the same random variable.

```
> dev.from.mean <- qmq(x, rep(mu, d))
> dev.from.mean.plus.two <- qpq(dev.from.mean, rep("2", d))
> dev.from.mean.minus.two <- qmq(dev.from.mean, rep("2", d))
> hrep <- addHin(as.numeric(qsign(dev.from.mean.plus.two) < 0),
+     "1/2", hrep)
> hrep <- addHin(as.numeric(qsign(dev.from.mean.minus.two) > 0),
+     "1/2", hrep)
```

These inequality constraints require that the median of X be in the closed interval with endpoints $\mu - 2 = 3$ and $\mu + 2 = 7$.

We now want to add some constraints on the variance of X but do not know what is reasonable. Since any linear constraint is maximized or minimized at an extreme point of the constraint set, we check what are the extreme points of the current constraint set and what the variance of X is at each.

```
> vout <- scdd(hrep, incidence = TRUE)
> extremes <- vout$output[ , - c(1, 2)]
> squared.dev.from.mean <- qxq(dev.from.mean,
+     dev.from.mean)
> extreme.var <- qmatmult(extremes, cbind(squared.dev.from.mean))
> extreme.var <- as.vector(extreme.var)
> extreme.var[order(q2d(extreme.var))]
```

```
 [1] "0"    "1"    "2"    "2"    "3"    "3"    "4"    "4"    "4"
[10] "6"    "6"    "8"    "8"    "9"    "19/2" "19/2" "10"   "10"
[19] "21/2" "21/2" "11"   "11"   "23/2" "23/2" "16"
```

```
> more.extreme.var <- extreme.var[sapply(vout$incidence,
+     function(foo) nrow(hrep) %in% foo)]
> more.extreme.var[order(q2d(more.extreme.var))]
```

```
[1] "9"    "19/2" "10"   "21/2" "11"   "23/2" "16"
```

The points which are rows of the matrix **extremes** are the extreme points of the current constraint set (represented by **hrep**). The components of the vector **extreme.var** are the values of the variance of the random variable X at those extreme points. The components of the vector **more.extreme.var** are the values for the subset of those points at which the constraint determined by the last row of **hrep** is binding (holds with equality). This is the upper bound median constraint, so this makes the point $\mu + 2 = 7$ a median of the random variable X.

Now we find the upper and lower quartiles of those variance values.

```
> fred <- q2d(more.extreme.var)
> sally <- quantile(fred, type = 1)[c(2, 4)]
> varbound <- more.extreme.var[fred %in% sally]
> varbound <- unique(varbound)
> varbound <- varbound[order(q2d(varbound))]
> varbound

[1] "19/2" "23/2"
```

Finally, we constrain the variance of X to be between those bounds.

```
> hrep <- addHin(qneg(squared.dev.from.mean),
+     qneg(varbound[1]), hrep)
> hrep <- addHin(squared.dev.from.mean,
+     varbound[2], hrep)
```

This completes the construction of the constraint set. The R object **hrep** specifies it.

4 Faces

This is a fairly complicated convex polytope.

```
> fout <- allfaces(hrep)
> dims <- unlist(fout$dimension)
> length(dims)

[1] 1103

> f <- tabulate(dims + 1, nbins = d + 1)
> f

 [1]   46 178 308 305 185   67   13    1    0    0
```

There is one (improper) face of dimension $d-2$, which is the whole constraint set. The dimension is $d-2$ because there are two equality constraints: the components of λ sum to one and the mean of X is μ. There is also one (improper) face which is the empty set (and which is not counted above, since the function **allfaces** only lists nonempty faces). There are 46 vertices (faces consisting of a single point), 178 edges (faces which are line segments), 13 facets (proper faces of maximal dimension, in this case $d-3$), and 67 ridges (faces of dimension one less than the dimension of facets, in this case $d-4$).

5 Simulation Truth

In order that our simulated data be a good test case, we want the inequality constraints to have effect. For this reason we choose the simulation truth parameter value (the λ we use to simulate multinomial data) to satisfy two of the four complicated inequality constraints with equality (where "complicated" means not the nonnegativity constraints, which are all satisfied with strict inequality). We choose the two upper bound constraints, which are the last row and the third to last row of `hrep`.

We make the simulation truth parameter value a relative interior point of the ridge satisfying these two upper bound constraints with equality as well as the equality constraints.

```
> ncons <- nrow(hrep)
> inies <- sapply(fout$active.set,
+     function(foo) all(c(ncons - 2, ncons) %in% foo))
> inies.max.dim <- max(unlist(fout$dimension[inies]))
> inies.of.max.dim <- inies &
+     unlist(fout$dimension) == inies.max.dim
> lambda.truth <- fout$relative.interior.point[inies.of.max.dim]
> length(lambda.truth) == 1

[1] TRUE

> lambda.truth <- lambda.truth[[1]]
> lambda.truth <- q2d(lambda.truth)
> lambda.truth

[1] 0.37946429 0.02008929 0.02008929 0.02008929 0.02008929 0.02008929
[7] 0.02008929 0.42187500 0.07812500
```

Clearly this λ satisfies the nonnegativity constraints. We check "by hand" using the most obvious code, that it satisfies the other constraints.

```
> x <- q2d(x)
> mu <- q2d(mu)
> sum(lambda.truth)

[1] 1

> sum(lambda.truth * x)

[1] 5

> sum(lambda.truth[x < mu - 2])

[1] 0.3995536

> sum(lambda.truth[x > mu + 2])

[1] 0.5

> sum(lambda.truth * (x - mu)^2)
```

```
[1] 11.5
```

```
> q2d(varbound)
```

```
[1]  9.5 11.5
```

We see that, indeed, all of the constraints are satisfied at this point, and two of the complicated inequality constraints are satisfied with equality, and, of course, this is all that are possible to satisfy with equality, because the median cannot simultaneously be at both its lower and upper bound and similarly for the variance.

6 Data

Make up data.

```
> y <- rmultinom(1, size = n, prob = lambda.truth)
> y <- as.vector(y)
> if (any(y <= 0)) stop("must have all y values strictly positive")
> y
```

```
[1] 366  22  21  23  16  22  23 426  81
```

7 Remark

None of the work done to this point in this document would need to be done for a real data analysis. The data vector y would be given (it would be the data). The equality and inequality constraints would be determined by prior knowledge (perhaps influenced by other data about the random variable X that is involved in the complicated constraints). A real data analysis would start here.

8 Affine Hull

We determine the affine hull of the constraint set.

```
> equalities <- hrep[ , 1] == "1"
> equalities
```

```
 [1]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
[12] FALSE FALSE FALSE FALSE
```

```
> av <- scdd(hrep[equalities, ])$output
> av.point <- av[av[ , 1] == "0", ]
> av.lines <- av[av[ , 1] == "1", ]
> av.point <- as.vector(av.point[- c(1, 2)])
> av.lines <- av.lines[ , - c(1, 2)]
> av.point
```

```
[1] "-3" "4"  "0"  "0"  "0"  "0"  "0"  "0"  "0"
```

```
> av.lines

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] "1"  "-2" "1"  "0"  "0"  "0"  "0"  "0"  "0"
[2,] "2"  "-3" "0"  "1"  "0"  "0"  "0"  "0"  "0"
[3,] "3"  "-4" "0"  "0"  "1"  "0"  "0"  "0"  "0"
[4,] "4"  "-5" "0"  "0"  "0"  "1"  "0"  "0"  "0"
[5,] "5"  "-6" "0"  "0"  "0"  "0"  "1"  "0"  "0"
[6,] "6"  "-7" "0"  "0"  "0"  "0"  "0"  "1"  "0"
[7,] "7"  "-8" "0"  "0"  "0"  "0"  "0"  "0"  "1"
```

The vector `av.point` is a point in the affine hull of the constraint set. The rows of the matrix `av.lines` are a basis for the tangent space of the affine hull of the constraint set (the vector space parallel to it).

We now change notation to follow Remark 6 of the paper.

```
> m <- t(q2d(av.lines))
> lambda.zero <- q2d(av.point)
```

Our `m` is M in the paper, and our `lambda.zero` is λ_0 in the paper.

We check to see that these objects behave well with respect to the constraints (as they must if the work above is correct). First we check that λ_0 satisfies the equality constraints (it does not satisfy the inequality constraints, and does not need to).

```
> sum(lambda.zero)

[1] 1

> sum(x * lambda.zero)

[1] 5
```

Second we check that the columns of M satisfy the equality constraints with the right-hand side changed to zero (because movement along these vectors should go from λ_0 to other points satisfying the equality constraints).

```
> as.vector(rbind(rep(1, d)) %*% m)

[1] 0 0 0 0 0 0 0

> as.vector(rbind(x) %*% m)

[1] 0 0 0 0 0 0 0
```

9 Point Estimates

```
> alpha.hat <- y
> lambda.hat <- alpha.hat / n
```

In notation of Section 4.2 of the paper, we are using the improper prior determined by setting the hyperparameter ξ to the zero vector. Our `alpha.hat` is $\hat{\alpha}_n$ in the paper, and our `lambda.hat` is $\hat{\lambda}_n$ in the paper.

10 Beta

Now we find the vector (3) and the matrix (4) that are the mean and variance of the normal distribution for the new parameter β that lives on the affine hull.

First, we find β_n^* given by (3).

```
> lout <- lm(lambda.hat - lambda.zero ~ 0 + m,
+     weights = 1 / lambda.hat)
> beta.star <- lout$coefficients
```

We can also do this by literally implementing (3).

```
> beta.star.too <- solve(t(m) %*% diag(1 / lambda.hat) %*% m) %*%
+     t(m) %*% diag(1 / lambda.hat) %*% cbind(lambda.hat - lambda.zero)
> all.equal(as.vector(beta.star), as.vector(beta.star.too))

[1] TRUE
```

Second, we find the variance matrix given by (4).

```
> varmat <- solve(t(m) %*% diag(1 / lambda.hat) %*% m) / n
```

11 Simulate Equality Constrained Posterior

```
> library(MASS)
> nboot <- 1e5
> beta <- mvrnorm(nboot, beta.star, varmat)
> lambda <- beta %*% t(m)
> lambda <- sweep(lambda, 2, lambda.zero, "+")
```

This code simulates β vectors (each row of the matrix `beta` is one such simulation. Then we transform to λ using (2), except since `beta` is actually a matrix, must use sweep instead of add.

Each row of the matrix `lambda` is a simulated normal random vector having the required mean vector and variance matrix. We check that all of these simulations satisfy the equality constraints.

```
> range(lambda %*% cbind(x))

[1] 5 5

> range(apply(lambda, 1, sum))

[1] 1 1
```

12 Apply Inequality Constraints to Simulation

Now we apply the inequality constraints to the simulated realizations of the unconstrained posterior. To do this we need the constraints in the form $B\lambda \leq a$.

```
> inequalities <- hrep[ , 1] == "0"
> bmat <- q2d(hrep[inequalities, ])
> avec <- as.vector(bmat[ , 2])
> bmat <- bmat[ , - c(1, 2)]
> bmat <- (- bmat)
```

This code makes matrix `bmat` and vector `avec`, which are B and a in mathematical notation, that determine the inequality constraints as described.

Since `lambda` is actually a matrix, we must sweep instead of add

```
> goodies <- lambda %*% t(bmat)
> goodies <- sweep(goodies, 2, avec)
> goodies <- apply(goodies < 0, 1, all)
> mean(goodies)
```

```
[1] 0.53917
```

Our Monte Carlo (MC) sample is the "goodies."

```
> lambda <- lambda[goodies, ]
```

Now we check — the dumb way to be safe — that our MC sample does indeed satisfy the constraints.

```
> dim(lambda)
```

```
[1] 53917     9
```

```
> range(apply(lambda, 1, min))
```

```
[1] 0.001703099 0.025639433
```

```
> range(apply(lambda, 1, sum))
```

```
[1] 1 1
```

```
> range(apply(lambda, 1, function(lambda) sum(lambda[x < mu - 2])))
```

```
[1] 0.3776154 0.4096745
```

```
> range(apply(lambda, 1, function(lambda) sum(lambda[x > mu + 2])))
```

```
[1] 0.4676482 0.4999998
```

```
> range(apply(lambda, 1, function(lambda)
+     sum(q2d(squared.dev.from.mean) * lambda)))
```

```
[1] 10.96291 11.50000
```

```
> q2d(varbound)
```

```
[1]  9.5 11.5
```

13 Remark

That ends the simulation. The rows of the matrix **lambda** are independent and identically distributed Monte Carlo simulations of the (approximate, large sample, asymptotic) constrained posterior distribution.

14 Importance Weights

Suppose we were to use the asymptotic approximation as an importance sampling distribution to sample the correct (finite n) distribution.

The Dirichlet posterior distribution has unnormalized probability density function

$$g(\lambda) = \prod_{i=1}^{n} \lambda_i^{\hat{\alpha}_i - 1} \tag{5}$$

and the asymptotic approximation has unnormalized probability density function (1). Their ratio is the unnormalized importance weights.

```
> log.g.lambda <- apply(lambda, 1, function(lambda)
+     sum((alpha.hat - 1) * log(lambda)))
> log.h.lambda <- apply(lambda, 1, function(lambda) {
+     foo <- lambda - lambda.hat
+     bar <- sum(foo^2 / lambda.hat)
+     return(- n * bar / 2)})
> weigh <- log.g.lambda - log.h.lambda
```

The vector **weigh** now contains the log unnormalized importance weights. Because of the constraints, we do not know the normalizing constants for the distributions. Hence these unnormalized importance weights are useless for importance sampling. By normalizing them, however, we do make them useful (Geyer, 2011, Section 11.1).

```
> weigh <- weigh - max(weigh)
> weigh <- exp(weigh)
> weigh <- weigh / sum(weigh)
```

If w_i are these importance weights (components of our R vector **weigh**) and λ_i are the samples (rows of our R matrix **lambda**), then for any function f the weighted average

$$\sum_{i=1}^{m} w_i f(\lambda_i)$$

is an unbiased estimator of the true posterior expectation $E\{f(\lambda) \mid y\}$, assuming this posterior expectation exists (that is, that f is integrable with respect to the posterior distribution).

It will be a good estimator so long as the importance weights are not too uneven. So let us look at them. The following R statements make the plot Figure 1.

```
> par(mar = c(5, 4, 1, 1) + 0.1)
> plot(log.h.lambda, weigh, log = "y")
```

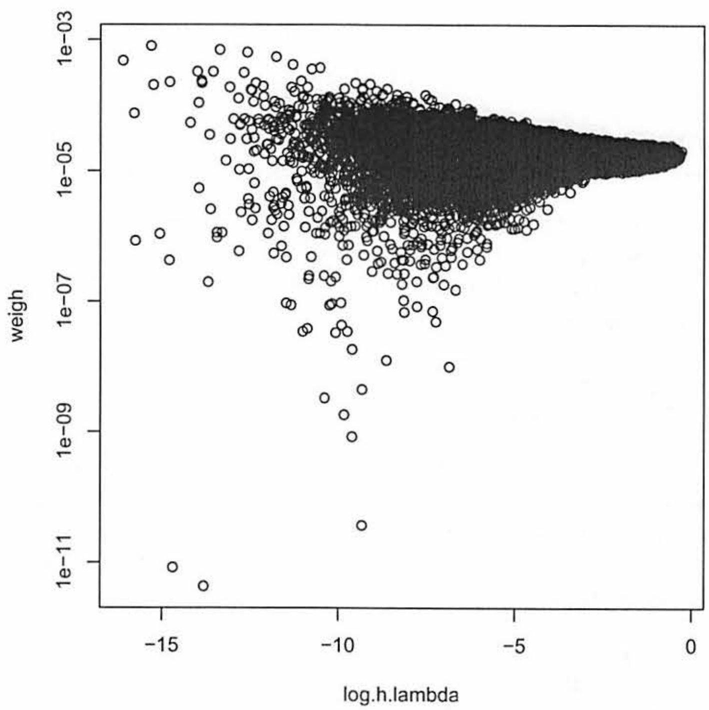

Figure 1: Normalized Importance Weights versus Log Unnormalized Density of Asymptotic Normal Distribution (vertical axis is log scale).

We see from the figure that the normalized importance weights are not too unevenly distributed. One indication of this is Figure 1. Another is the ratio of the maximum importance weight to the average.

```
> max(weigh) * length(weigh)

[1] 42.84962
```

Hence we conclude that importance sampling will work well (for these simulated data y), and we can use our asymptotic approximation to calculate expectations with respect to the exact posterior distribution.

15 Discussion

It is somewhat surprising that we went to all that trouble to simulate the constrained normal approximation to the constrained exact posterior and then didn't use the constrained normal approximation except as an importance sampling distribution. But this does make sense. We need the asymptotic approximation because we do not know how to simulate an equality constrained Dirichlet distribution (we do know how to simulate an unconstrained Dirichlet distribution because it is a product of beta conditionals and marginals, but that is no help), whereas we do know how to simulate an equality constrained normal distribution. Thus the usefulness of the normal approximation is that it allows us to impose the equality constraints. This cannot be done by rejection sampling because the equality constrained distribution lies in a lower-dimensional affine subspace that has probability zero under the unconstrained distribution. In contrast the inequality constraints can be imposed (as we did above) by rejection sampling (simply reject the points that do not satisfy the inequality constraints).

This importance sampling scheme will work well when the normal approximation is good (when n is large) and will not work well when it isn't (when n is small). In this respect, it is no different from any other use of asymptotic approximation.

Rather than produce a massive simulation study (repeat the above with lots of different random number generator seeds, different sample sizes `n` and different dimensions `d`), we have produced a "simulation schema" that allows interested readers to redo this document with different choices of these quantities. Industrious readers can with a bit more work also change the constraints.

References

Geyer, C. J. (2011). Importance sampling, simulated tempering, and umbrella sampling. In *Handbook of Markov Chain Monte Carlo*. Edited by S. P. Brooks, A. E. Gelman, G. L. Jones, and X. L. Meng. Chapman & Hall/CRC, Boca Raton, pp. 295–311.

Geyer, C. J. and Meeden, G. D. (2009). R package `rcdd` (C Double Description for R). Current version 1.1-3 (14 Jul 2009). `http://www.stat.umn.edu/geyer/rcdd/`.

Geyer, C. J. and Meeden, G. D. (submitted). Asymptotics for constrained Dirichlet distributions. Revised and resubmitted to *Bayesian Analysis*.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

BIOMETRICS 27, 201–11
March 1971

The following paper, although couched in humorous terms, makes, we believe, a number of serious points. Readers are invited to respond, not necessarily in kind, by a Letter to the Editor.

THE PRACTICAL PSYCHOLOGY OF BIOSTATISTICAL CONSULTATION[1]

LYON HYAMS

Division of Biostatistics, Department of Community Medicine, Rutgers Medical School, Rutgers The State University, New Brunswick, New Jersey 08903, U. S. A.

SUMMARY

Beginning with the premise that consultations with biological scientists are frequently characterized by communication difficulties, this paper tries to provide insights into their etiology through the consideration of the seemingly different expectations and behaviors of consultants and clients. General issues and interpersonal problems are brought into focus by stereotypic characterizations. Suggestions for upgrading the consulting relationship are advanced that depend on the empathetic understanding of the client's position and a more realistic self-appraisal.

It takes a certain amount of 'chutspah' to lecture on this theme to a readership of whom many have had more experience in consultation than I have had. I can partially justify the attempt, at least to myself, by pointing to my joint background in medicine and statistics, which is helpful in considering the bilateral problems encountered in consultation. Also, if we can accept the proposition that consultation is often characterized by difficulty, then attempts to deal with this 'trouble' rationally are worthwhile. Other authors recently have considered problems of consultation in several excellent papers listed in the references. These articles are organized along different lines and, for the most part, emphasize different aspects of consultancy problems.

THE IDEAL CONSULTATION

Let us first consider the 'Ideal Consultation' as fantasized by the statistician in order to illustrate his strivings and expectations. To qualify a consultation as ideal is to deny its empirical meaning. The 'Ideal Consultation' is *not* a consultation. It is a working-together, a voluntary meeting of minds and union of energies whose prime aim is to seek a 'truth.' In such meetings both parties are familiar with each other's basic language. The biologist has had a few courses in basic statistics and thus recognizes statistics as a unique and valuable discipline. The statistician has also done his homework and has familiarized himself with the names and the relationships of the fauna in the experimentor's jungle. Since knowledge and understanding breed sympathy

[1] This paper represents an extension and restructuring of remarks presented in a Letter to the Editor, *Biometrics 25*, 431–4, 1969.

and respect, the researcher esteems the statistician as an expert representative of this most important science. His appreciation for the statistician's unique contribution grows by leaps and bounds with the experience of his individual talents. Needless to say, the feeling is mutual. Meetings are stimulating; they are productive in thought and in product. The work forms a gestalt (the whole is greater than the sum of its parts). The research poses challenging statistical problems that are fun to work at: the sort of thing that keeps one busy at a scratch pad during supper while the wife silently suffers (or throws a fit). In unhurried time the deliberations proceed to a design, an experiment, and an analysis that confirms everyone's best hopes. The (multiple) reports are easy to write. Sometimes the biologist's name is first, sometimes the methodologist's; it hardly matters. These manuscripts are received enthusiastically by journal editors and their 'expert' reviewers don't give the team a hard time. First experiments lead naturally to others and the information generated finds a significant practical application. Ultimately it saves human lives or curtails misery. Finally, but justly, the co-workers are awarded the Nobel Prize in Medicine and quite naturally donate their stipends to schools encouraging interdisciplinary approaches to problem solving.

This dream may touch upon the desires and expectations of the consulting biostatistician. It illustrates his utopian strivings, yearnings towards the three R's: Reality, Respect, and Reward.

But what of the realities? The following negative client stereotypes (some of whom have stepped into my office) can illustrate the dirty backyard of these experiences. While I acknowledge that most clients are reasonable men, these extreme characterizations can help clarify the pathodynamic mechanisms present in the more routine interactions.

CLIENT STEREOTYPES

The *Probabilist's* only interest is a significant 'p' value. And the lower the number is, the happier he feels. Men who have not smiled in twenty years will break out in spontaneous laughter when hearing that $p < 0.001$. While most individuals of this species will know that this is 'significant,' it is doubtful whether they will understand the meaning of the Type I error. Less erudite investigators, not quite so sure of what to do with this number, may ask, their voices tinged with sarcasm, 'What are the magical words to use with this thing?' This is a moment of truth. Responses to such a question separate the men from the boys. Finally the Probabilist comes in two subvarieties. Those who leave the form of the test up to you, and those who insist on a particular, inevitably inappropriate one.

The *Numbers Collector* began his work three years ago last May. He has performed a kaleidoscopic array of N^N experiments and stumbles into your office trying to balance eight vertical inches of smudged data sheets clutched to his chest. Settled in a chair, he has the greatest difficulty in explaining his experiments and framing his questions. He reeks a defeatist aura of 'too complicated for anyone but me to understand, much less explain.' Nevertheless, he has a specific request. In essence, he says: 'Here is a solid basalt mountain and somewhere in them rocks are precious jewels, randomly deposited. Mine the jewels!'. He also makes it quite clear that it is *very* important to finish the job within the next three days because of grant renewal and publication deadlines. And so the biostatistician might dig in. Find the ruby; the experimenter will sell it and pocket all the cash. Find nothing but colored glass and the consultant is the culprit.

The *Sporadic Leech* is the casual acquaintance who stops the consultant in the hall or interrupts his lunch for a brief friendly chat. Starting with the weather or the world situation, he quickly directs the conversation to its real purpose: the analysis of his experiment or 'What do you think I should do with the following data?'. Accepting all suggestions in a seemingly casual fashion, he will make it clear that the biostatisticians expertise is a nice, but unnecessary verification of something he already knows. All suggestions but one, that is. It is a mistake to recommend an office visit. He doesn't want the invitation and will interpret it as a sign of the consultant's incompetence. For above all, he is determined not to make a formal visit and will never acknowledge his needs for guidance to himself or anybody else. He is the 'do-it-yourselfer, come-what-may.' The consultant can have a playful interchange with him for months, or years, but one fateful day he will receive a serious phone call from him requesting literature references for the analyses performed and incorporated in the final manuscript. And if the statistician has been suckered in this far, he can hardly choose this time to be cagey. Fortunately this form of client never mentions the consultant's name—not even in a footnote—so the only thing one has to lose is his local reputation.

The *Amateur Statistician* believes that the really intelligent people are good politicians, psychologists, and statisticians; all without formal training or experience. The only reason he consults at all is because he can't spare the six or seven hours to master an advanced text in mathematical statistics. The consultant's role in the ensuing drama is that of the technician. The client plays the director, quick in calling the type of analysis to apply or the test to use, overwhelming one's protestations and pleas for thinking time with a conspiratorial and knowing smile, saying in effect: 'Now don't make more of it than it is'; as if thought before analysis was a gambit in the statistician's confidence game. This common sense devotee, seemingly worldly wise and thick skinned, is really ultrasensitive. Don't tell him that he is wrong. He'll never believe it.

The *Long Distance Runner* was born and bred in the slums of the southeast Bronx or in some debilitated midwestern village. He is insecure and determined to 'make it.' At the tender age of three and a half years, he started to run fast and hard. Forty years have passed. When you meet him (always at his office) it is obvious that he is a man of stature. Magnanimous in his prosperity, he welcomes the consultant with a warm smile and a firm handshake and spends the following five minutes telling him about the exciting future and significance of his work. But nothing about the work itself. The interview quickly ends. An arm over the consultant's shoulder, he escorts him to his second or third in command for the dirty details, setting himself to receive an important phone call from Russia. While obviously he has no time for you (or his work), he is charming and pays well. His real order as conveyed by his staff emerges slowly: 'Find something to say—anything!' The critical question is: 'Why is this man still running?' He has excellent reasons. Two steps behind him, scrambling as fast and furiously as is he—arms outstretched, threatening to engulf and pull him down, down, down into pit, are a troupe of horrible demons, his Mistakes!

The description of these stereotypes is useful for a number of purposes. Foremost among these, it gives the author a much needed catharsis. It may vicariously provide the reader with a similar service. But apart from these egocentric concerns, extreme characterizations help in defining and studying problems. They sharpen our focus from vague dissatisfaction to indisputable specific concerns. Granting it will be rare to deal with the extreme case, outlier analysis can be used profitably in detecting and understanding similar trends in the more routine consultation. There is no line between extreme and routine. We can better understand 'normal' functioning through the examination of the abnormal.

NEGATIVE CHARACTERISTICS OF STEREOTYPES

One characteristic applies to all of the described types: while these investigators may be looking for the truth, they are not seeking it with the statistician. His approach to reality is not generally asked—nor wanted. Consequently, it is easy for the consultant to suspect erroneously that such investigators miss the motivational core of all honest scientific efforts: the search for what 'actually is.' This interpretation can occasion strong feelings of disappointment and bitterness.

All of these stereotypes clearly lack an understanding and appreciation of statistical thought as a distinct and valuable discipline. In their confusion they tend to underestimate or overestimate the operational domain of biostatistics. Some, like the Amateur Statistician and the Probabilist, equate the science of statistics with a sometimes necessary and sometimes evil tool, a minor offshoot of Common Sense or Algebra that journal editors and the statistical lobby foster on them. Feeling duped, it is easy for the clients to transfer these hostile feelings from the subject of statistics to its representative. Just as easily, the consultant may interpret the obvious lack of respect for the science as one for himself. In addition, such an underestimation of the complexity and value of statistics often puts the consultant in the uncomfortable position of supplying too simple and probably wrong solutions to the problems posed.

Another group of clients, like the Numbers Collector or the Long Distance Runner, must feel that the statistician is a voodoo priest who will bring order to chaos and clarity to confusion, both without the effective cooperation of the biological worker. Such persons, knowing little about the discipline, have preferred to project magical powers and extent to it in order to serve their own needs. Consequently, they expect the biostatistician to solve their problems for them and thus shift an unrealistic amount of responsibility onto his shoulders. Apparently, human beings faced with a piece of the unknown will behave in these unnatural and dichotomous ways: irrationally limiting and extending the domain of its influence. It is this fact, that our science is essentially unknown to them, that forms one of our central problems.

Another negative characteristic that working with such investigators presents is that these consultations can leave the consultant with a feeling of being used. The feeling is perfectly legitimate. He is being used. This stimulates a hostile reaction in him, related to the insult implied and to the feelings of helplessness it engenders. And even if this 'use' is a mutual one—the consultant may use the investigator as a data source, an hour statistic, or an easy publication source—the compromise does not result in an optimal interaction. This form of cold war, competing with one another to get as much as each possibly can, while giving as little as possible, denies both parties access to the adventure of mutual exploration and the comfort of the comradeship of striving together towards a common goal. It denies the truth-seeking aspirations. Such interactions, characterized by an absence of openness, warmth, and honesty, do not contribute to a superior product.

There are many other disadvantages attached to working with clients as described. For example, the consultant's contributions are not often directly rewarded with coauthorship on subsequent publications. The demands of the work do not usually present interesting mathematical problems. Taken altogether, it is difficult to achieve the three R's: Reality, Respect, and Reward.

Finally, some meetings seem to be forced. Some clients, for diverse reasons, would rather not ask for the cooperation of the biostatistician. Nevertheless, they believe they can't do what is required themselves. Such meetings contain all the qualities of interactions which would lead to bad relations and mutual avoidance in outside life. This occurs frequently in our profession as well.

Are there solutions to these problems? I believe so. We can begin by examining the options available to the consultant cursed with a case load of many of these negative types. (This approach is a natural extension of the development of the extreme case. While most clients are a sympathetic, intelligent, and appreciative lot, a study of their typology does not contribute greatly to the understanding of our problems.)

AVAILABLE OPTIONS

The consultant can withdraw from painful contacts, 'act out' hastily towards the client, or accept his limited role. These solutions to problems in diplomacy are not optimum. It would be more productive to improve the unsatisfactory aspects of his interactions though this choice presents the greatest difficulties.

The ability to improve consultations is dependent upon a number of factors. These include: The acceptance of the early negative components of the meeting, a constant awareness of one's own motivation and expectations (some aspects of which have been discussed), and a better understanding of the client. This last point must be elaborated.

THE VIEWPOINT OF THE CLIENT

Very often the statistical consultant is placed—or places himself—in the role of the professional critic versus the creator-client. The research worker has labored arduously, long, and often imaginatively to synthesize his product. His work, whatever the quality, nevertheless magically represents an extension of his personal value. Naturally, he is sensitive to hearing about its (his) imperfections. Yet concomitantly he must seek this critique in order to better his work. And it is the consultant—critic who has the unpleasant responsibility of telling him what is wrong with his baby. The client, while often objectively understanding this form of constructive disintegration, is nevertheless pained on an irrational but important emotional level. It may color his other dealings with the consultant or manifest itself in an increasing intolerance to any criticism. In addition, the client's natural sensitivity may be heightened by an undue critical emphasis arising from 'acting-out' biostatisticians.

The client's need for the consultant's understanding and appreciation of the special qualities of his world is closely allied to the above concerns. His chosen problem leans on a structural base of biological knowledge and present practicalities which must be grasped, at least in their essentials, in order for meaningful exchange to occur. Imagine the dismay and frustration the client must feel at having to explain background to a consultant who doesn't know his basic language. How can he possibly feel that his labor and his contributions to the subtleties of the problem are understood if the consultant is unfamiliar with the definitions, assumptions, and factual base of the working domain? Further, unless the biological problems are grasped by the statistician, he may produce a biologically meaningless but beautifully intricate statistical analysis which can only increase the communication gap. In total, the investigator can be left with the feeling that he and the subject matter he represents are not understood, that the consultant speaks a different language for which translation is desperately needed.

The difficulty in communication becomes more acute when submitting manuscripts to medical journals which do not favor mathematical sounds beyond a t test. While the statistician is giving him methodological advice, the biologist is simultaneously trying to translate the 'statisticese' into written medicine, while thinking of the present editorial policy and the audience this journal reaches. Also unless one appreciates the relative import of the various components of the investigator's present work, the statistician can make the mistake of thinking that the problem presented to him represents the total research effort. He may then logically but erroneously conclude that all matters stand or fall on a competent solution. Arguing from this viewpoint takes on a more directed and aggressive aspect than if the problem presented were one sidelight of many statements which jointly made a contribution to a general physiological hypothesis. Such difficulties in communication leave the client feeling shaky and he may not want to continue the relationship.

The biological worker also may fear or show a disinclination for contacts with a discipline he knows little to nothing about. While these avoiding reactions may be related to the embarrassment of revealing personal information gaps, it is more likely that the 'fear of the strange' is, in this case, attached to a different psychodynamic mechanism. The client may be primarily concerned (though unaware) that 'this statistics business' is really vital for his work but simultaneously that it is beyond his ability to master. This is true either by virtue of the difficulty of the concepts or the lack of time for study. In a sense he is faced with the impossible choice of committing himself to a new and difficult study or subjecting his work to a method beyond his comprehension and control. Quite naturally he finds both paths repellent and he experiences himself positioned on the edge of the blade, wishing it would all go away.

Another possible point of contention between the client and the consultant relates to differences in personality. While some purists will state that personality type is a soft variable, most any good psychologist will say that the statistician tends to orderly structures and preciseness (if not obsessiveness) while the biologist prefers the rambling, global approach to glory. Physicians, as a special group, present additional difficulties. They are prone to excessive independence and aggressiveness, they are impatient with scientific frills, and they would rather run the show than work in a team. The practicing statistician, on the other hand, specializes in team work. Such differences must occasion some conflict. In addition, Alvan Feinstein has suggested several types of statisticians who may brutalize the client. A brief description of these stereotypes—of which characteristic fragments can be found in many—is justified.

CONSULTANT STEREOTYPES

1. The *Model Builder* fits any and every data problem set to a model he is presently interested in or knows something about. It matters not whether he investigates the questions that are being asked by the client or those that are biologically important. For that matter, this type isn't really interested in hearing the client's story. He had posed his own a priori questions before the client knew him. The Model Builder is like the drunkard looking for his lost key under the street lamp although he dropped it in the dark alley. He justifies his search by pointing out that there is light in the place he is looking.

2. *The Hunter* is the statistician counterpart of the Numbers Collector who directs you to 'mine the mountain.' The Hunter will subject every data set to an exhaustive and extensive computer analysis. For a relatively simple problem with scanty data he will ultimately present the investigator with 14 vertical inches of print-out, containing 17 significant results. These numbers do not bear a relationship to anything on the face of this earth except themselves. While the client may initially accept these authoritative materials with reverence, it will not take him long to figure out that he has a bag of wind.

3. *The Gong* is a consultant who starts every conference by drawing a bell shaped curve.

4. *The Traditionalist* is convinced that nothing really important has hap-

pened in statistics since R. A. Fisher, and consequently limits himself to a restrictive working vocabulary. He views computers as the devil's work.

5. *The Randomophiliac* firmly believes that it doesn't matter what else you do, as long as you've 'randomized' well. He is like the mother who catches her 14-year old daughter in a sexually compromising situation and admonishes her by saying 'as long as you don't smoke, honey.'

6. *The Quantophreniac's* position is: It doesn't matter if you observe what you want to as long as you can get a hard measurement.

7. *The More Data Yeller* (he needs no further description).

8. *The Nit Picker* will always focus his attention on the inconsequential but debatable. He will enlarge minor issues out of reasonable perspective and quickly reduce a real and tremendous contribution to a potentially horrendous error in reality testing. (My manuscripts are usually reviewed by this type).

Thus it is no wonder that the client may approach the interaction with ambivalence and hesitation. He does not want his creation criticized, he fears—and feels—a lack of understanding and respect both for his field and for himself, he is hesitant to become involved in a discipline he knows little about, and he may not receive an appropriate treatment of his problems. He suffers also from a lack of Reality, Recognition, and Reward—problems with which practicing statisticians are familiar.

REMEDIES IN THE INTERPERSONAL DOMAIN

This discussion suggests certain remedial activities that the consultant can undertake to improve his interactions and to better communication.

To offset his role as a critic, he can begin by expressing his appreciation and preferably his enthusiasm for the effort and work performed. To be able to do this honestly, it is clear that the consultant must spend his first energies in understanding the biological problems and the practical difficulties associated with their solution. He should try to see the world of the investigator through his eyes—phenomenologically. This necessitates a certain amount of openness, biological knowledge, and an initial empathetic expectation. I believe that only if he leaves himself open to this new view—the other person's view—he can appreciate most honest efforts in research and data gathering. However, this may necessitate postponing solution-giving to the following meetings. He should develop the concept of an on-going relationship with the investigator, who will be able to appreciate and accept it if it responds to his world and interest. I have never heard complaints from clients while they are discussing their progress and ideas with a receptive, uncritical audience. If the clinical statistician first assumes the mantle of the 'receptive listener,' the biologist must, in turn, be more sympathetic to the consultant's comments. The inveterate Gong or the Model Builder will never reach this stage of enlightenment. They are in verbal or written action before a first-level understanding is reached.

Another active goal which the consultant must determine to achieve is the subtle education of the client. This learning process should not overwhelm him in its complexity for this would reinforce his defensive rejection of statistics. Rather, the simple logic and clarity of statistical methods should be stressed and applied to his work. For example, a Type II error should be simply illustrated if it is pertinent to the client's results: that is, if he has found no differences and there is a lot riding on the negative finding. The consultant should not offer solutions that are beyond the comprehension of the experimenter or his ability to describe them unless it is specifically agreed that the client will perform backup at meetings and interviews. Six t tests are better than one ANOVA to the researcher who cannot encompass ANOVA. This example is especially pertinent since multiple t testing is standard practice in the medical and biological journals; ANOVA is seen only sporadically.

It is also good practice to avoid presenting too much, too soon. At an early stage of interaction the client will find it hard to accept so many new, vital truths that no one else possesses. While this is often the truth, it is asking too much for the outsider to appreciate it. Slow education and support should be supplied continuously from the first meeting through the final manuscript. If the experimenter trips over the statistics, he won't ask for help again. Attempts at educating the client can also be made in groups at a center of affiliation. Voluntary courses can turn out to be a simple method of introducing basic ideas to the faculty and 'drumming up trade.' The wary client is also given a chance to look the statistician and his subject over before he commits his ego and energies. Sagacious attempts at statistical education will hopefully improve the tone of these consultations and result in more interesting mathematical stimulation for the consultant. They might even lead to requests to design experiments—usually a late development.

In accepting the negative components of initial interactions while actively trying to overcome them, the consultant seems to be doing more than the investigator and perhaps this provides him with a reason for feeling resentful. One logical retort is to point out that since the consultant can control only his own behavior, he has no other choice. A more productive reply, however, involves the consideration of the help-orientation of the client. Like a patient, he may have a child-like expectation of his authority figure, the consultant. In asking for help he doesn't want to see in him anything *but* what he wants and needs. Personal foibles, human weaknesses? These are for other people. He doesn't need disagreement; that is not what he came for. He came for help and in this orientation finds it more difficult to evoke in himself the additional energies necessary to meet the consultant even halfway. People in need rarely consider the needs of their helper. This is difficult to accept because of the professional status of the client. The challenge and art of consultation are to transform his raw need to appreciation and influence. These remarks seem to imply that it is good starting policy to give the client what he requests by providing the service he wants (if it is at all reasonable to do so). By so doing it is difficult to push him away. Maybe all one can do for the Probabilist the first time around is to give him that significant probability value. But he might return for something better.

It is also obvious that the time available to the consultant will determine attitudes and actions toward clients with negative characteristics. A statistician securely intrenched in some institutional structure may have enough requests for consultations to be able to restrict his case load to the easiest or most interesting problems. The fledgling is forced to take all comers and to upgrade their statistics in order to improve his status. In this sense, it is the less experienced statistician who plays the more important role with the difficult client. Potentially he can win us a few more friends. The 'Big Shot' works with the ordained, the saved. The Fledgling often works with the fringe of potential converts. The Big Shot has usually lost the missionary zeal, a spirit desperately needed. It is one that generally brings unrecognized truths to the unaware in order to help them and it implies a certain reluctance of the subjects to being helped. One certainly must have a measure of missionary zeal to accomplish this end with clients with many negative characteristics.

Finally, if the biostatistical consultant spends more than half of his time consulting, he is in too precarious and weak a position to function efficiently. The good statistical consultant must negotiate with his clients from a position of strength, not weakness, a position of understanding, not need. In this way he will better be able to demand reasonable rewards.

CLOSURE

While I have tried to confine my remarks to the interpersonal domain, it is obvious that educational and institutional problems contribute to individual difficulties. I hope to consider the structural milieu, its problems and remedies, in a future article. However, I cannot help but point with great satisfaction to the formation of the ASA Committee 'Teaching of Statistics in the Medical Sciences,' whose general goal of upgrading didactic methods acknowledges the need of a self-critical reappraisal. More specifically, one subcommittee, CEOMSI (Committee to Effect the Optimization of Medical Statistical Interaction), composed of both physicians and statisticians offers the promise of the long overdue productive dialogue. I hope that other groups: the public health statisticians, the computer men, biomathematicians, biomedical engineers, etc., will undertake similar ventures. Finally we must also consider fusion of these mathematically oriented approaches to the biological truths. In so doing, we can reap our greatest rewards.

LA PSYCHOLOGIE PRATIQUE DE LA CONSULTATION EN BIOSTATISTIQUE

RESUME

En partant de la constatation que les consultations avec les chercheurs en biologie sont souvent caractérisées par des difficultés de communication, cet article essaie d'éclairer leur étiologie en partant du fait qu'apparaissent différents chez les consultants et les clients les résultats attendus et le comportement.

Des problèmes de fond et des problèmes entre personnes sont mis en évidence par des attitudes stéréotypiques. On avance des suggestions pour améliorer les relations au cours de la consultation; elles dépendent de la compréhension de la position du elient et d'une connaissance plus exacte de soi-même.

REFERENCES

Cameron, J. M. [1969]. The statistical consultant in a scientific laboratory. *Technometrics* *11*, 247-54.

Cox, C. P. [1968]. Some observations on the teaching of statistical consultancy. *Biometrics* *24*, 789–802.
Daniel, C. [1969]. Some general remarks on consulting in statistics. *Technometrics 11*, 241–6.
Sprent, P. [1970]. Some problems of statistical consultancy. *J. R. Statist. Soc. A 133*, 139–65.